ESTUDIO SOBRE LA IDONEIDAD DE LA RAZÓN DE PRECIPITACIONES

José Antonio López Díaz[1]
[1]*Agencia Estatal de Meteorología*
jlopezd@aemet.es

RESUMEN
Se aborda una evaluación de la idoneidad del modelo de razón de precipitaciones entre estaciones próximas comúnmente usado para relleno de lagunas y estudios de homogeneidad. La idoneidad de la adecuación a la razón se valora por la suma de las distancias ortogonales cuadráticas a la recta de regresión ortogonal adecuadamente normalizadas.
Se presentan los resultados de un análisis por meses del año del índice de desviación de la razón introducido, clasificando también vertientes y cuencas hídricas. Se hace un estudio de la significación estadística de los resultados basado en una simulación por bootstrap de los valores del índice de desviación de la razón con pares de series de precipitación de observatorios aleatorios.

Palabras clave: razón de precipitación, relleno lagunas, regresión ortogonal

ABSTRACT
The ratio model between the precipitations of two close observatories commonly used for gap filling and homogeneity analysis is assessed. The squared distance sum to the orthogonal regression line is used in order to measure the suitability of the ratio model. The results of an analysis by month and watershed of the deviation from the ratio index that is introduced are shown. A significance study of these results is also carried out, based on a bootstrap simulation of the deviation from ratio index between pairs of observatories to obtain critical values for the statistical test.

Key words: precipitation ratio, gap filling, orthogonal regression

1. INTRODUCCIÓN

Un modelo muy usado en climatología para la relación entre precipitaciones de observatorios próximos es el modelo de la razón simple. Dadas las precipitaciones de un par de observatorios próximos en un mes dado x_i e y_i este modelo postula:

$$y_i = A*x_i , \qquad i= 1, 2, \ldots,N \qquad (1),$$

siendo N el número de años con datos y A la razón.
Este modelo está en la base del método tradicional muy empleado de estimación de lagunas de la precipitación en un observatorio a partir de observatorios próximos de

la Razón Normal (Chow *et al*, 1988). Así mismo, en el contexto del estudio de la homogeneidad y la corrección de inhomogeneidades en series de precipitación, el conocido test SNHT introducido en la climatología por Alexandersson (1986), se basa en el modelo (1) también. El SNHT supuso en su día un avance importante en el tratamiento de la homogeneidad en series de precipitación, y continúa siendo utilizado él mismo o alguna de sus variantes mejoradas como el RHTestV4 (Wang 1997).
Para valorar estadísticamente la adecuación de la precipitación mensual en España al modelo (1) se han seleccionado pares de observatorios suficientemente próximos y con datos síncronos suficientes. En la sección de metodología se presenta una justificación del índice de desviación aplicado para testear el modelo (1), así como de la técnica de bootstrap empleada para generar el nivel crítico del test.

1.1 Datos

Para tomar un número grande de parejas de observatorios próximos con series suficientemente largas de datos síncronos mensuales de precipitación, se ha partido de los ficheros de existencias de dato mensual de precipitación del Banco Nacional de Datos Climatológicos de la AEMET. Se ha seleccionado el periodo 1951-2000 que a priori dispone de muchas series de precipitación. Se ha generado en primer lugar la matriz de dimensiones (11817, 50, 12) con 0/1 en cada posición según haya o no dato del mes y año del periodo 1951-2000 en el observatorio respectivo del total de 11817. A continuación se generó la matriz de distancias entre cada pareja de observatorios, y se seleccionaron las parejas de observatorios a menos de 5 km de distancia. Resultaron 19014 parejas de entre las que se seleccionaron aquellas que, en al menos un mes, tienen 20 años con dato síncrono, con lo que se redujeron las parejas a 1825.

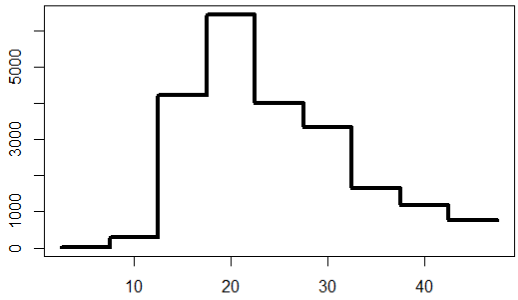

Fig. 1: Histograma de frecuencias de nº de años de dato síncrono en todas las series mensuales para todas las parejas de observatorios seleccionadas.

En la figura 1 se muestra cómo queda la distribución del número de años con dato síncrono en las 12 series mensuales de todas las parejas seleccionadas, es decir, el número de series que entran en consideración es igual al número de observatorios seleccionados por 12. La figura indica que en general predominan los pares de meses con un número de años síncronos con dato superior a 20, de hecho son un 86.2% de los pares.

2. MÉTODOS

En esta sección se aborda la definición de un índice que valore el grado de desacuerdo de los datos con el modelo de la razón simple para lo que se establecen en primer lugar varios requisitos para el mismo. La última parte se dedica al estudio de la significación estadística del índice definido.

2.1. El índice de desviación de la razón IDR: requisitos y definición

Buscamos definir un índice que valore el grado de desacuerdo de los pares de precipitaciones (x_i, y_i) con el modelo (1), y que de forma convencional varíe entre 0 y 100, de tal forma que los valores pequeños indiquen mayor validez del modelo. A un tal índice lo denominaremos índice de desviación de la razón IDR. Podemos a priori sentar los siguientes requisitos deseables para un IDR:

C.1) Debe ser invariante de escala, es decir, independiente de la unidad de medida de la precipitación.

C.2) Su valor no debe modificarse al permutar los papeles de los dos observatorios, ya que el modelo (1) se expresa igualmente bien con las x's en el primer miembro y las y's en el segundo sin más que sustituir A por 1/A.

C.3) Los pares de valores de precipitaciones (0, 0) (o muy próximos) si se dan (caso habitual en verano en muchos observatorios), por adecuarse perfectamente al modelo (1) para cualquier razón A, deben hacer que el IDR decrezca, de forma monótona al ir añadiendo pares nulos a una serie.

Para construir un IDR parece natural partir del ajuste óptimo de la razón A, el IDR medirá de alguna forma la discrepancia de los pares de valores respecto de la recta del plano $y=A*x$ con el A del ajuste. Este ajuste a la vista de C.2) no puede basarse en la regresión lineal ordinaria sin término independiente, ya que esta postula un modelo no simétrico en x e y. Una alternativa consiste en la regresión ortogonal, empleada en la fundamentación del paquete de homogeneización CLIMATOL del software libre R (Guijarro Pastor, 2013). Esta regresión busca minimizar la suma de distancias cuadráticas de los puntos del plano (x, y) a una recta, es decir, medidas según la perpendicular a la recta que pasa por el punto, y no como la regresión lineal, que busca minimizar la suma de las distancias cuadráticas según el eje Y. De acuerdo a esta definición la regresión ortogonal trata de forma simétrica a los dos ejes X e Y y satisface pues C.2).

Una vez encontrada la recta de la forma (1) óptima según la regresión ortogonal, denotemos al valor mínimo de la suma de las distancias cuadráticas de los puntos a la recta por SCOr. Está claro que cuanto mayor es SCOr peor es la adecuación del modelo de razón a los datos, por tanto es un buen punto de partida para el IDR.

Para satisfacer la invariancia de escala C.1) necesitamos normalizar la SCOr. Una posibilidad es partir de que la suma de distancias cuadráticas de unos puntos cualesquiera a una recta es siempre menor o igual que la suma de distancias cuadráticas de los mismos puntos a cualquier punto de la recta, y en concreto al punto de la recta que minimiza dicha suma de distancias cuadráticas. Esto se deduce inmediatamente de que la distancia cuadrática de un punto a uno de la recta se descompone, de acuerdo al teorema de Pitágoras, en la suma del cuadrado de la distancia a la recta, más el otro cateto sobre la recta al cuadrado. Si denotamos esta

segunda suma de distancias cuadráticas al punto minimal de la recta de regresión ortogonal por SCPunOr definimos:

$$IDR = 100 * SCOr/SCPunOr \quad (2)$$

2.2. Demostración de que el IDR satisface los requisitos

Veamos que el IDR definido por (2) satisface todos los requisitos. En primer lugar, puesto que SCOr<=SCPunOr, IDR estará en el rango [0, 100]. Como además todas las definiciones son geométricas es fácil ver que IDR cumple C.1), y por supuesto C.2) al basarse en la regresión ortogonal. Falta por verificar la propiedad C.3) de disminución monótona de IDR con la adición de pares (0, 0).

En primer lugar resulta claro de la definición de la recta de regresión ortogonal por el origen, que esta no varía al adicionar a los datos pares (0, 0), y que por tanto SCOr no varía tampoco (pues los pares (0, 0) están a distancia nula de cualquier recta por el origen). Por tanto C.3) estará probado si SCPunOr aumenta monótonamente con la adición de pares nulos.

Veamos que SCPunOr aumenta por dos razones: porque la contribución de los N puntos iniciales a SCPunOr crece al adicionar N0 pares (0, 0), y también porque los N0 pares (0, 0) contribuyen positivamente a la nueva SCPunOr. Lo segundo es evidente ya que una distancia cuadrática siempre es positiva. En cuanto a lo primero, partimos del teorema que afirma que la suma de distancias cuadráticas de un conjunto de N puntos a un punto **P** del plano, denotada por SCD(**P**), cumple:

$$SCD(\mathbf{P}) = SCD(\mathbf{G}) + N * |\mathbf{G}\,\mathbf{P}|^2 \quad (3)$$

siendo **G** el centro de masas de los puntos cuyas coordenadas son las medias de las coordenadas X e Y de los puntos, y **G P** es el vector que une **G** y **P**. La igualdad (3) se puede demostrar por ejemplo descomponiendo SCD(**P**) en las contribuciones de las x's y las y's, y aplicando el correspondiente conocido teorema para números cualesquiera a ambas coordenadas.

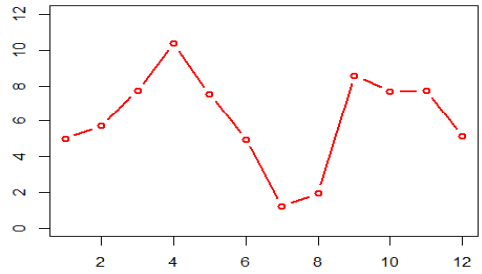

Fig. 2: Valores críticos al 10% para el IDR para cada mes del año

Aplicando (3) deducimos que el punto **R1** de la recta de regresión ortogonal que minimiza la SCD(**R1**) para los N puntos iniciales tiene que estar en la perpendicular

bajada desde su centro de masas **G1** a la recta. Además, de nuevo por (3), la SCPunOr de estos puntos iniciales vale:

$$\text{SCPunOr1} = \text{SCD}(\mathbf{R1}) = \text{SCD}(\mathbf{G1}) + \ + \text{N}^* \ |\mathbf{G1\ R1}|^2$$

Cuando adicionamos N0 pares nulos, el centro de masas **G2** del nuevo conjunto de N+N0 puntos está en la recta **O G1**, siendo **O** el origen del plano, pero más cerca de **O** que **G1**. En cualquier caso, salvo que **G1** ya fuera **O** (caso trivial de ajuste perfecto, pues al ser las precipitaciones necesariamente positivas esto implica que los N puntos iniciales ya eran (0,0) todos), **G2** no coincidirá con **G1**. En consecuencia el nuevo punto de la recta de regresión ortogonal **R2** que minimiza la SCD(**R2**) para los N + N0 puntos, obtenido bajando desde **G2** por la perpendicular a la recta de regresión ortogonal, que sigue siendo la misma, no coincidirá con **R1**. Por tanto la nueva suma de SCPunOr para los N puntos iniciales, SCPunOr2, valdrá en virtud de (3):

$$\text{SCPunOr2} = \text{SCD}(\mathbf{G1}) + \ + \text{N}^* \ |\mathbf{G1\ R2}|^2 \ > \ \text{SCPunOr1}$$

pues $|\mathbf{G1R1}|^2 < |\mathbf{G1R2}|^2$ por la definición de **R1**. Con esto queda demostrado que el IDR dado por (2) satisface todos los requisitos.

2.3. Significación estadística del IDR

Dado un conjunto de datos (x, y), planteemos un contraste de hipótesis, con hipótesis nula consistente en que el valor de su IDR es aleatorio, frente a la hipótesis alternativa consistente en que los datos se ajustan al modelo de razón simple (1). De acuerdo a la definición de IDR, deberemos rechazar la hipótesis nula para valores suficientemente bajos del IDR.

Puesto que la distribución teórica del IDR en condiciones de aleatoriedad es desconocida, e incluso es difícil precisar qué tipo de aleatoriedad tiene sentido en este caso, se ha aplicado una técnica de bootstrap no paramétrico (Efron *et al*, 1994) para encontrar los valores críticos del IDR para diferentes niveles de significación. Para ello se han tomado, para cada mes

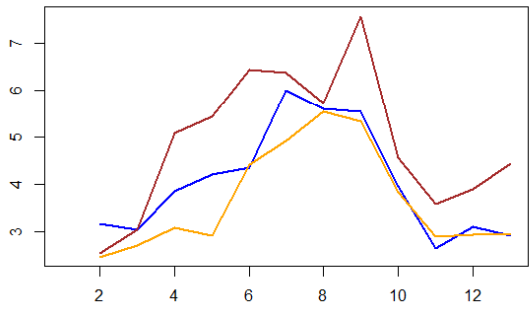

Fig. 3: Mediana del IDR cada mes en abscisas: azul para vertiente atlántica, naranja para vertiente mediterránea y marrón para Canarias.

del año, pares aleatorios de observatorios de entre el conjunto todos los observatorios, pero sin mantener su emparejamiento de tal forma que la distancia entre ellos varía también aleatoriamente sin estar restringida a < 5 km. Se ha impuesto luego la condición de que en al menos un mes cada par aleatorio tenga al menos 20 datos síncronos. De esta forma se han producido para cada mes 10000 valores aleatorios del IDR. En la tabla figura 2 se muestran los valores críticos al 10% del IDR para cada mes, es decir, los percentiles 10 de cada serie de valores aleatorios del IDR, ya que debemos rechazar la hipótesis nula para valores bajos del IDR.

Se puede apreciar en la figura 2 que el valor crítico del IDR muestra un ciclo estacional claro, con valores más altos en primavera y otoño. Los valores especialmente bajos en verano pueden tener que ver con la presencia de pares de datos de precipitación nula o muy baja en los dos observatorios, lo que sabemos por C.3) que debe bajar el valor del IDR. Por otra parte la importante variación del valor crítico entre meses hace que a priori haya dos posibilidades con resultados diferentes, según que apliquemos a cada mes el valor crítico del mes, o bien apliquemos a todos los meses el mismo valor crítico global. Este último se obtiene hallando el percentil 10 del conjunto de todos los datos de precipitación mensual aleatorios, incluyendo todos los meses, y resulta ser de 5,62.

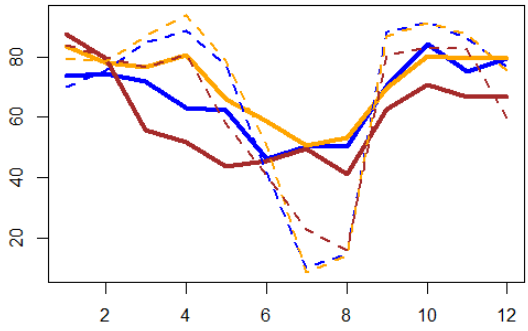

Fig. 4: Porcentaje de pares de observatorios con un IDR significativo al 10% por meses en abscisas y vertientes hidrográficas según color como en la figura 3. Líneas continuas con nivel crítico global, líneas a trazos nivel crítico de cada mes.

Aunque es cierto que el valor crítico de cada mes tiene en cuenta la particular climatología, no está claro que sea este el más adecuado, pues el modelo de razón (1) especifica una simple relación, y no está del todo claro cuál es el conjunto aleatorio que forma la base de la hipótesis nula de este test. En este trabajo se empleará, salvo en los primeros resultados que se comentan más abajo, el único nivel global sin distinción de meses, pues parece a priori que el nivel por meses penaliza en exceso la adecuación al modelo en los meses veraniegos. Esto puede deberse a que en verano en España, aunque los observatorios estén alejados, las parejas de valores próximas a cero abundan mucho, y por tanto no se genera una variedad efectiva suficiente de casos que debe estar en la base de la variación aleatoria bajo la hipótesis nula. Es decir,

los resultados del verano con el nivel por meses no serían muy representativos por falta de representatividad del valor crítico obtenido.

3. RESULTADOS

Se presentan a continuación los resultados de un análisis de la significación del modelo de razón por meses y cuencas hídricas en primer lugar, y después se investiga la posible influencia de la altitud de los observatorios y de la cantidad de precipitación.

3.1 Estudio por meses y cuencas hídricas

En la figura 3 se han representado las curvas de evolución mensual, para las vertientes mediterránea, atlántica y Canarias, de la mediana del IDR para los pares de observatorios próximos objeto de este estudio. Las tres curvas muestran una variación a lo largo del año similar, con máximos en los meses más secos centrales del año. En Canarias el IDR toma valores en general superiores a las otras vertientes de la Península y Baleares, y las vertientes mediterránea y atlántica muestran valores similares de la mediana del IDR, un poco superiores en la atlántica.

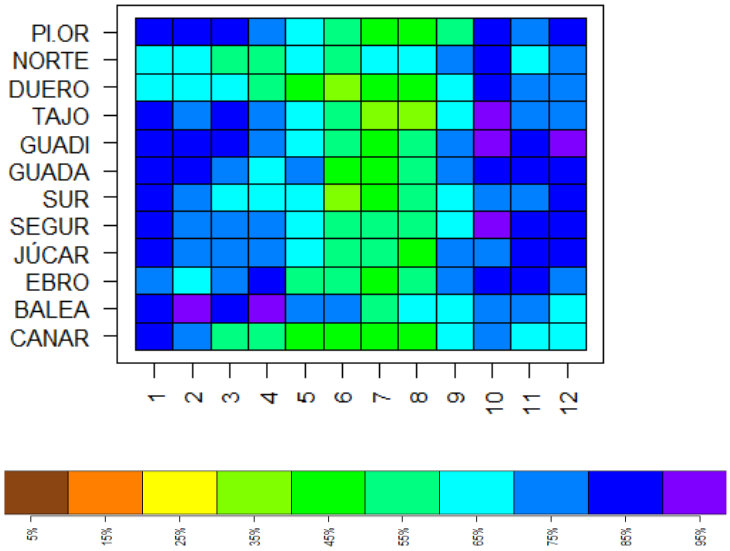

Fig. 5: Para cada mes y cuenca hidrográfica, porcentaje de pares de observatorios con adecuación del modelo de razón significativa al 10%.

En la figura 4 se muestra el resultado de aplicar el test sobre la significación estadística del IDR descrito en 2.3 a un nivel de significación del 10%. Cada porcentaje de la figura corresponde pues al de pares de observatorios, en la vertiente y mes respectivos, para los que el IDR es significativamente bajo y por tanto el modelo de razón simple (1) es pertinente según este test. Se han representado tanto los porcentajes de

significación aplicando el nivel crítico global (líneas continuas) como aplicando el nivel crítico propio de cada mes (líneas a trazos).
En línea con los valores de la mediana del IDR, el modelo de la razón se adecúa peor en las Canarias, seguido de vertiente atlántica y de vertiente mediterránea, con mayores diferencias en la primera mitad del año. Con nivel crítico global, los porcentajes de significación del modelo oscilan entre 45% y 85%, con peor adecuación en verano.
Un desglose de estos porcentajes de significación del modelo de razón por cuencas hídricas se muestra en la figura 5, en la que se ha aplicado el nivel crítico del IDR global, como en el resto del trabajo. En general las diferencias entre cuencas son mayores en la primera mitad del año, en la que las cuencas norte y Duero son las que se ajustan peor al modelo de la razón, y Baleares la que muestra mejores valores.

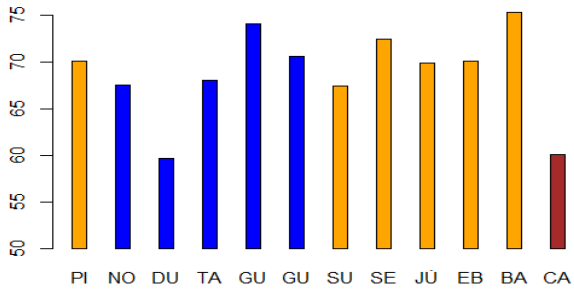

Fig. 6: Para cada cuenca hidrográfica, porcentaje medio anual de pares de observatorios con adecuación del modelo de razón significativa al 10%.

En la figura 6 se han reflejado los valores medios de los porcentajes de significación del modelo de la razón por cuencas (en el mismo orden que en la figura 5). La cuenca del Duero es la de peor adecuación del modelo, junto con Canarias, mientras que Baleares y Guadiana destacan por el buen ajuste del modelo. Con todo, las diferencias entre cuencas son algo menos importante que las diferencias entre meses.

3.2 Dependencia con la altitud y la precipitación media
En la figura 7 se analiza la posible dependencia de la adecuación de la precipitación a la razón con la altitud de los pares de observatorios, separando en dos clases los observatorios según que su altitud sea superior o no a la media. Se aprecia que los observatorios bajos muestran mejor adecuación todos los meses, pero es en los meses de la primera mitad del año en los que las diferencias en el porcentaje de adecuación significativa al 10% de nivel de significación son mayores, alrededor del 15% en abril, mayo y junio.

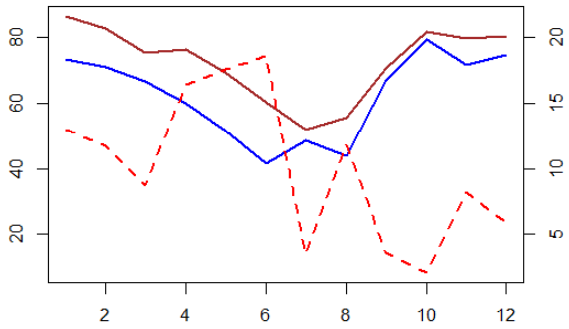

Fig. 7: Porcentaje de pares con un IDR significativo al 10% por meses en abscisas, marrón pares de observatorios con altitud media inferior a la mediana, azul altitud media superior a la mediana. La diferencia entre las dos curvas anteriores en línea de trazos, con escala a la derecha.

La posible dependencia de la adecuación del modelo de razón simple con el total de precipitación se visualiza en la figura 8, en que el desglose de los observatorios se hace según la mediana de la precipitación. Resulta evidente que la dependencia es pequeña en comparación con la dependencia con la altitud, y las diferencias en porcentaje de pares de observatorios con adecuación significativa al 10% entre los dos grupos son inferiores en valor absoluto al 5% en todos los meses, con cambio de signo a lo largo del año.

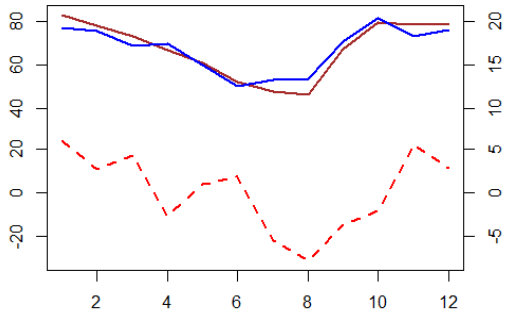

Fig. 8: Como en la fig. 5, pero la curva marrón se refiere a pares de observatorios con precipitación media inferior a la mediana del mes, azul precipitación media superior a la mediana.

4. DISCUSIÓN

Se ha definido un índice para medir la desviación de un conjunto de pares de precipitación (mensual) el modelo de razón simple (1) para el que se ha probado que satisface una serie de requisitos que parecen muy razonables a priori, y que limitan

considerablemente las posibilidades. Para este índice se ha definido un test estadístico con valores críticos obtenidos por una técnica de bootstrap.
Los resultados del grado de significación estadística del modelo de razón simple aplicando este índice IDR muestran una importante variación inter-mensual para las dos vertientes peninsulares y las Canarias, con la peor adecuación del modelo en los meses veraniegos. Entre las dos vertientes peninsulares hay una pequeña ventaja de la vertiente mediterránea sobre la atlántica en los meses primeros del año, y Canarias muestra peor adecuación que las dos vertientes peninsulares. El análisis por cuencas hídricas incluyendo todos los meses indica que la cuenca del Duero y Canarias son las de menor adecuación de la razón, y las Baleares y las del Guadiana las que mejor se adaptan.
Se ha explorado también la posible dependencia con la altitud del observatorio, obteniéndose cierta ventaja en la adecuación a la razón en los observatorios más bajos, sobre todo en la primera mitad del año. En cambio no se ha encontrado dependencia con el total de precipitación.

AGRADECIMIENTOS

Deseo expresar mi agradecimiento a mi colega de la AEMET José Antonio Guijarro, que me llamó la atención sobre esta cuestión y con el que mantuve interesantes intercambios de ideas.

REFERENCIAS

Alexandersson, H. (1986). A homogeneity test applied to precipitation data, J. of Climatology, Vol. 6, 661-675.

Chow, V.T., Maidment, D.R., Mays, L.W. (1988). Applied Hydrology, Mc Graw Hill Book Company, ISBN 0-’07-010810-2

Efron, B., Tibshirani, R. (1994). An Introduction to the Bootstrap. Chapman & Hall/CRC. ISBN 978-0-412-04231-7.

Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. Biometrika. 68 (3): 589–599.

Guijarro Pastor, J.A. (2013) Users Guide to Climatol. Recuperado de http://www.climatol.eu/index.html

Wang, X. L., Wen, Q. H., Wu H. (2007). Penalized maximal t test for detecting undocumented mean change in climate data series. J. Appl. Meteor. Climatol. 46 (No. 6), 916-931. DOI:10.1175/JAM2504.1